Autonomy Technology Overview

Autonomy Whitepaper

Index

Autonomy Technology Overview

A plethora of information exists throughout an enterprise, created and stored in different formats across multiple technologies, and exponentially growing by the moment. This information is the lifeblood of any organization as it is used to communicate to customers, power mission-critical operations, and drive strategic business decisions. As this information continues to grow, organizations require technology that is able to unify information scattered across an organization, make sense of the data, and instantly connect people with the relevant information that is required to drive business success.

Autonomy's Intelligent Data Operating Layer is a powerful infrastructure technology that surpasses keyword-matching to not only retrieve, but understand the meaning of all information that exists throughout an organization. Built on adaptive pattern recognition technology and probabilistic modeling, IDOL forms a conceptual and contextual understanding of digital content, and extracts meaning from the manner in which people interact with that content. Supported by the ability to recognize over 1,000 different file formats and securely connect to over 400 repositories, the technology provides advanced and accurate retrieval of the valuable knowledge and business intelligence that already exists within the organization. From a single platform, companies can access and process any piece data in any form, including unstructured data such as text, email, web, voice, or video files, regardless of its location or language.

IDOL exceeds the scope of legacy approaches to deliver unprecedented speed, scalability, visibility, and sophisticataed analysis and decision-making capabilities. Using advanced functions automatically and in real time, only IDOL can perform keyword and conceptual search, while eliminating the traditionally manual and costly operation of processing and analyzing information. Not only does IDOL connect people with relevant, personalized, and accurate information; it also recognizes patterns hidden within disparate pieces of information to identify emerging trends and lucrative business opportunities. The undisputed power of IDOL lays it its ability to learn and adapt to new and evolving technologies and information, providing businesses the agility to respond to and stay ahead of emerging business requirements. Today, 20,000 global companies, law firms, and federal agencies as well as 400 OEMs leverage Autonomy's technology as their information processing layer.

Autonomy's Vision

Autonomy was founded upon a vision to dramatically change the way in which we interact with information and computers, ensuring that computers map to our world, rather than the other way around. Human-friendly or unstructured information is not naturally found in the rows and columns of a database, but in documents, Web pages, presentations, videos, phone conversations, emails and IMs. We are facing an increasing deluge of unstructured information, with 80% now falling into this category and, according to Gartner, the volume of this data doubles every month. As the amount of unstructured information multiplies, the challenge for the modern enterprise is trying to understand and extract the value that lies within this vast sea of data, whilst minimizing the risk. Many companies believe that access to information is the answer to dealing with the unstoppable spread of information of all forms – if people can find information, they can process it themselves. Autonomy believes that although access to information is important, there is far greater value in forming an understanding of data and automatically processing it, freeing up people to focus on higher-value activities that computers are unable to do. By providing a pan-enterprise software infrastructure that automates advanced operations, Autonomy presents customers with a compelling value proposition. With this ability, Autonomy enables organizations to penetrate their information silos, derive maximum value from their corporate assets, and boost productivity while minimizing the risks endemic to information proliferation.

Meaning Based Computing (MBC)

Autonomy is the acknowledged leader in the rapidly growing area of MBC. MBC refers to the ability to form an understanding of all information, whether structured, semi-structured or unstructured, and recognize the relationships that exist within it. This allows computers to harness the full richness of human information, bringing meaning to all data, regardless of what or where it is. Through sophisticated functionality and analytics, MBC automates manual operations in real-time to offer true business value. MBC extends far beyond traditional methods such as keyword search which simply allow users to find and retrieve data. Keyword search engines, for example, cannot comprehend the meaning of information, so they only find documents in which a specific word occurs. Unfortunately, this inability to understand information means that other documents that discuss the same idea (i.e. are relevant) but use different words, are often overlooked. Equally, documents with a meaning entirely different to that which the user searches for are frequently returned, forcing the user to alter their query to accommodate the search engine. While Autonomy offers and acknowledges the importance of keyword technologies, MBC enables advanced capabilities based on a conceptual and contextual understanding of information. Some of the key functionalities of MBC such as automatic hyperlinking and clustering are simply not available in keyword search engines. For example, automatic hyperlinking, which connects users to a range of pertinent documents, services or products that are contextually linked to the original text, requires that the meaning of the original document is fully understood. Similarly, for computers to automatically collect, analyze and organize information, computers have to be able to extract meaning. Only MBC systems can do this.

"Autonomy is now the only remaining pure play vendor in Enterprise Search space"

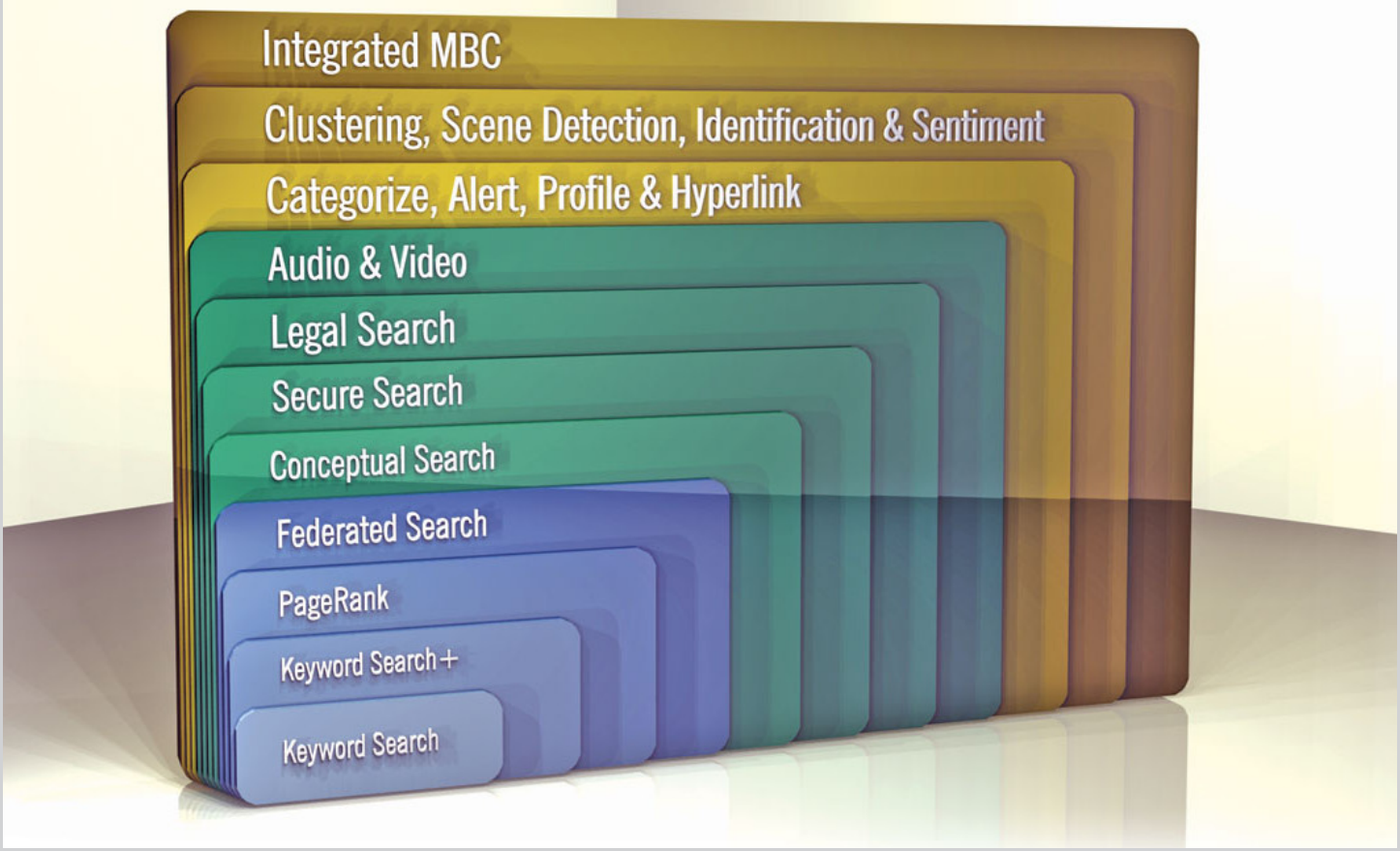
—Goldman Sachs, 2008

Evolution of Search

From the time of the very first computers, their inability to process human-friendly, “unstructured” information has posed a considerable challenge. The modern IT industry was founded on the principle that, for example, if the number in “Column 3” goes to zero, the computer will automatically order more stock for the warehouse – in other words, the position of a piece of information tells the computer what to do with it – and a tremendous amount of effort has been poured into sorting and distilling unstructured information into tidy rows and columns. Increasingly, structuring information in this way does not represent a viable solution, not only because of the incredible amount of manual effort required but because by organizing information in this way, its richness and subtleties are lost. Consequently, attention has turned to finding alternative, more intelligent solutions to the problem of unstructured information and the journey towards integrated MBC began...

Keyword Search

Because computers were unable to understand the meaning of information, the seemingly obvious alternative was to simply search it in order to locate any keywords relevant to the desired subject. The problem with this approach is that the computer has no way of identifying what a given keyword means, and therefore cannot process the information afterwards. For example, if a user types in the letters “D-O-G” the computer has no concept of what that word means; it will simply identify all of the documents which contain that combination of letters, which might produce a list of results thousands of pages long.



Keyword Search +

In order to improve the results from straight-forward keyword searches, the technique was enhanced by adding a series of arbitrary rules so that the most relevant results would appear at the top of the list. For example, if the search term appears in the title of a document, five points are added to that result, and if it appears three times within the document, one point. This works to a certain extent but the important issue is that there is still no understanding of what a "D-O-G" is, or does. In addition, the rules have to be modified manually and become very costly to maintain every time a subject develops.

PageRank

On the Internet there is a simple trick to get around this problem because in many cases, the most popular information is also the most relevant. The importance or popularity of a Web page is approximated by counting the number of other pages that are linked to it, and by how frequently those pages are viewed by other users. This works quite well on the Internet but in the enterprise it is doomed to failure. Firstly, there are no native links between information in the enterprise. Secondly, if a user happens to be an expert, perhaps in the field of gallium arsenide laser diodes, there may be no one else interested in the subject, but it is still imperative that they find relevant information.

Federated Search

As a result of new regulatory drivers such as the FRCP, enterprises need to be able to guarantee that a search has covered absolutely every piece of relevant information across potentially hundreds of different repositories throughout the enterprise. Most search engines are not actually capable of doing this so they ask the original repositories to perform the search - a process known as federated search. Federated search is often advertised as an asset. However, in fact it creates significant problems because it generates vast increases in network traffic. Every time the user enters a query, each and every repository has to do a search, so a repository that previously ran a search perhaps 0.01 times per user per day, starts to glow white-hot. More importantly, all of the results are searched using a different algorithm which means that all of their relevance rankings are different and incompatible when compiling a results list. In addition, most of the underlying search algorithms used in the repositories are not compliant with the new FRCP. Consequently, federated search is not compatible with a pan-enterprise platform. All of the approaches described up to this point fit squarely into the mid-enterprise search market. A technology which is limited to these capabilities is not suitable for a true pan-enterprise deployment, for reasons that will now become clear.

Conceptual Search

A critical leap forward came with the ability to actually "understand" the idea behind a given phrase, and retrieve information which is conceptually related, even when a particular keyword is not used. So for example, if the user types in the letters "D-O-G", a conceptual search engine will retrieve all the information conceptually related to but not confined to the word "D-O-G", perhaps information about a "hound" as well as "walks" and different breeds of dog, because it understands the idea represented by the word. This is incredibly powerful because critical information is often missed because users do not always use the same search terms.

Secure Search

Security is absolutely paramount to the enterprise and the challenge this poses is staggeringly complex, from protecting the enterprise's intellectual property from unauthorized access, to ensuring internal compliance with an ever-growing list of regulatory requirements. Most users are not permitted to view most documents or even be aware that they exist. Typically, around 1/1000 documents should be available to each user and access privileges must be specific to each of the myriad of underlying repositories in the enterprise. Achieving air-tight security without significant performance degradation is a considerable challenge.

Legal Search

In order to scale without impeding performance, some technologies fail to search each document in its entirety. This prevents users from retrieving valuable information and it exposes the enterprise to significant compliance risk. Such technologies begin to calculate the relevance of each document at indexing time; however, if at the beginning of the calculation a particular result appears to be irrelevant, the engine will stop calculating, effectively assuming the result is not relevant without reading all the way through. Consequently, a relevant snippet of information on the last page of a hundred page report could be overlooked and the legal consequences could be absolutely catastrophic. In fact, the company CEO could go to jail because the search failed to retrieve all of the information required by the court.

Audio and Video Search

The full potential of multimedia content is often not utilized due to the fact that it has traditionally taken considerable manual involvement to process. Consequently, intelligence lies dormant in resources such as recorded meetings, training videos and broadcast content. True Pan-Enterprise Search technology automatically captures, encodes and indexes television, video and audio content from any source and provides users with the ability to search this with pinpoint accuracy and treat rich media content in the same way as more traditional forms of information.

Categorize, Alert and Profile

When computers "understand" information, they can start to automatically process it and begin to bring information to the user rather than the other way round. For example, through forming an understanding, computers can automatically create taxonomies, alert users to new and relevant information in real-time or automatically profile an individual's interests based on what they read and write, offering them interesting information without the need to search or connect with similar people.

Clustering, Scene Detection, Speaker Identification and Sentiment Analysis

Understanding information allows computers to cluster information, identifying inherent themes or clusters of conceptually similar information. In addition, using this approach it is possible to detect irregularities in everyday scenes for security purposes, identify well-known speakers in broadcast media and analyze conversations to detect positive or negative sentiment.

Integrated Meaning Based Computing

In examining the different approaches to the challenge of unstructured information, it becomes clear that the solution does not boil down to plain search. It is only through understanding the meaning of ALL information that computers are able to automatically process it and provide users with the ability to handle and maximize the value of this rich resource. MBC addresses the full range of information challenges and consequently forms the central requirement of major enterprise deployments all over the world.

Pan-Enterprise Search

The value of enterprise search continues to increase as organizations are expanding their use of the technology from business operations into information risk management. This broadening use of enterprise search has created a need for a platform solution that addresses Pan-Enterprise Search.

Convergence of Information Management, Information Governance and eDiscovery

The explosion of electronic information is leading to a convergence of information management, information governance and eDiscovery as all three disciplines attempt to control the growing volume of content in the face of regulatory pressure. Whether the issue is determining how long to keep a document, how to preserve, collect, review and produce that document for litigation, or simply how to make it findable amongst millions of documents for enterprise search, there is an acute need for an infrastructure solution that enables repeatable business processes. The benefits of deploying a single platform solution cannot be overlooked as it addresses the following challenges that plague stovepipe approaches:

- **The value of information changes over time:** A seemingly negligible email is just one lawsuit away from holding indispensable value for the enterprise. If the email is determined to be relevant to the impending case, a generic retention and disposition schedule must be replaced by a placement of legal hold. The value of all documents in an enterprise is subject to such vacillations. It is this very malleable nature of a document's significance that necessitates such close integration between legal hold and records management. It will therefore be difficult to achieve consistent enforcement of legal holds and retention policy using multiple platforms.
- **All three disciplines deal with the same corpus of data:** A traditional approach requires that when an enterprise starts its eDiscovery process (that of identifying, culling and reviewing the relevant data), it must re-index all the enterprise content. But this is the very same information that has already been indexed by the enterprise search engine (assuming a solution that can access all enterprise repositories). This costly and time-consuming endeavor can be eliminated when employing a single platform solution whose index is used by enterprise search, information governance, and eDiscovery. Autonomy provides all three solutions with no need to move data.

Pan-Enterprise Search Requirements

Autonomy is the acknowledged leader in Pan-Enterprise Search. Pan-Enterprise Search infrastructure must be integrative and agnostic, crossing company divisions, geographic locations, vendors, products, applications, operating systems and languages. It should support search and eDiscovery across all repositories that contain electronically stored information (ESI), including email, IM, voice, video and text across operational systems, archives and media; stored in centralized corporate servers, fileshares, desktops and handheld devices. The information should be 'plumbed into' once and leveraged for many applications across the enterprise – to extract business value and protect from information risk. True Pan-Enterprise Search demands:

- **Access to ALL data sources and file types:** Structured, semi-structured and unstructured. 80% of information within the enterprise is now unstructured, consisting of text, audio and video. This must be processed for regulatory reasons and to harness its true value.
- **Language independence:** Enterprises today have operations across the globe conducting business in numerous languages.
- **Compatibility with all Operating Systems in the enterprise:** Pan-Enterprise Search platforms need to be completely neutral within the enterprise and be able to work with any operating system.
- **FRCP compliance:** render all relevant ESI discoverable, regardless of format or location. To be FRCP compliant, Pan-Enterprise Search platforms:
  - Need to search ALL repositories.
  - Cannot perform jump out – a sleight of hand used to feign performance where the search engine stops looking across an index as soon as it is believed a large enough group of results has been assembled.
  - Need to produce auditable results – hence ALL data needs to be searched fully.
  - Need to be able to pass results to a hold function – ensuring that relevant ESI is preserved, not altered in any way or deleted.
- **Distribution and Fault Tolerance:** For organizations that are geographically distributed, local replicas should be automatically created and utilized where possible. Remote copies should only be used when a local system fails, thereby building fault tolerance, the benefits of local performance and a reduction of resource overhead into a single, seamless service.
- **Load Balancing:** Data should automatically be replicated across multiple servers and user requests should be load balanced across these replicas, guaranteeing performance, reducing latency and improving user-experience.
- **Mirroring/Failover:** Automatically generated replicas should be used to provide a pool of servers. The primary resource should be automatically selected and the system should switch to secondary systems if it fails so that service continues uninterrupted.

Autonomy's IDOL: A Single Platform Solution

By deploying IDOL as a standard information access platform across the organization, all electronic information existing within the enterprise can be automatically indexed, retrieved, processed and managed based on its meaning. IDOL not only allows organizations to understand the meaning of information, it also allows organizations to understand people's interactions with that information. Whether it is an operational query for business information or an eDiscovery request, the requirements for both enterprise search and FRCP compliant legal search can be achieved with IDOL. By forming a conceptual and contextual understanding of all enterprise information, IDOL also automates real-time policy management, proactively identifying and classifying information across all operational systems and content archives for consistent application of records management. With Autonomy's single platform solution, organizations have the flexibility to leave sensitive data at its source (Manage-In-Place) or move to Autonomy's intelligent archives at any point in time. Once the information is indexed, analysis can be performed simultaneously across Autonomy's archives and operational systems for compliance and litigation, ensuring consistency of results.

Federation: Not an Answer to Pan-Enterprise Search

Many organizations have attempted to provide Pan-Enterprise Search by relying on a federated search infrastructure. While this may be an appealing quick fix, enterprises should be aware of the limitations of this approach.

1. By federating searches to native search engines, one assumes that these engines are capable of effectively searching their own data. Given that many repositories rely on out-dated, end-of-life search products for their native search, this is not a reasonable assumption to make. These search engines rarely search all information in the repository, instead using sleight of hand techniques to perpetuate the illusion of performance. This poses a significant compliance risk now that the FRCP mandates that all Electronically Stored Information (ESI) be made searchable.
2. Federation does not scale. Sending every query to all of the native search engines greatly increases not only network traffic, but also load on the repositories themselves. When this problem is combined with common legacy security techniques (such as unmapped methods), the network traffic demanded by even a modest intranet search can be crippling. Furthermore, the speed at which the results list is returned is dependent on the speed of the slowest engine.
3. Federation does not necessarily produce a single, coherent, relevance ranking. Each search engine to which a query is federated determines relevance ranking independently and each engine will return a separate list of results .Compiling these results into one coherent list is not trivial.
4. Federation introduces data privacy risks. Relying on different search technologies means also depending on their security protocols to remain current and respectful of enterprise policies. Using a mishmash of search engines introduces data privacy risks that can reveal sensitive personal as well as enterprise information assets.
5. Valuable business insight can be lost when multiple search engines are used to power the enterprise knowledge base. Since both the enterprise and the user information (e.g. content consumption following a search) is scattered across siloed engines, user profiles are formed with far less contextual information. Search results consequently lack sufficient personalization, making the search process less meaningful for the user.

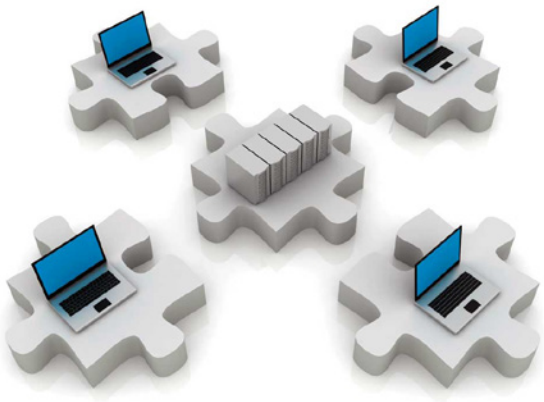
Recognizing the benefits of federation in certain situations, Autonomy can federate to any query-based system and intelligently interleave returning results. Autonomy adds intelligence to the traditional federation approach and compensates for its limitations in several ways:

- Allows search criteria to be pre-processed and results post-processed in nearly any way imaginable.
- Intelligently selects the specialist most likely to return a query result, avoiding the need to bombard multiple sources with the same query.
- Automatically merges results from different search engines into one coherent set.

Unifying All Enterprise Information

As a pioneer in the Meaning Based Computing (MBC) movement, Autonomy is a recognized leader in solving the difficult problem of analyzing unstructured content, which comprises 80% of enterprise data. But what about the rest of data in business applications, where a substantial amount of enterprise intelligence resides? Due to the challenge of processing different types of content in a siloed environment, most enterprise search vendors leave the processing of structured data to Business Intelligence (BI) technologies. However, maximizing business intelligence requires the proper integration and combined analysis of these disparate content types. For the heterogeneous enterprise that holds many sources of data, Autonomy's mature connector framework supports over 400 repositories to enable search across the entire enterprise corpus from a single interface. This allows for an unprecedented picture of the organization's information assets in one view. Autonomy supports structured data with the same level of intelligence, flexibility and precision as it does unstructured content, preserving complex relationships and automatically correlating relevant content by extracting concepts and entities from all data types.

Support for Structured Data



IDOL natively offers an array of search, navigation and analytics tools to process structured data. The advanced user can make SQL queries (FIND, JOIN, ORDER BY, SELECT, etc.) to pinpoint and manipulate search results, but IDOL offers far more sophisticated retrieval capabilities using an accessible, human-friendly query language. IDOL provides the same large number of operators to search structured data as it does to unstructured text – conceptual, keyword, search per field, conditional (i.e. give me results for products > $100K, cost < $15K, company name starting with "B"), wildcard, exact phrase, proximity, fuzzy search, relational and intersected taxonomy-based search, predictive spelling, stemming and synonym expansion, thesaurus query, search clustering, query-by-example, BIAS and many more. One example of an IDOL operator that is not offered in standard database search is BIAS, which allows the user to easily modify the relevancy calculation by giving more weight to a specified field. Given a movie database with title, director, lead actors, year, genre and short description as the fields, the user can choose to bias the "genre" field or the "director" field to determine the similarity between movies.

Difference from OLAP cubes

IDOL was designed from inception to absorb any data type, be it voice, video or text, and manipulate, compare and relate these objects based on a mathematical abstraction of the meaning within each source. This permits IDOL to augment human decision-making through self-discovery of relevant dimensions within the data sets. Such self-discovery can be constrained and controlled at the atomic level by human operators, but dramatically separates IDOL from the manually-dependent models of traditional OLAP and other legacy data management systems. IDOL can thus normalize and automatically relate items within heterogeneous and previously unorganized data sets, no matter if the ideas expressed reside in tables, emails, telephone calls or videos, based on a mathematical understanding of the concepts within it. Thus, unlike the OLAP model, the storage of the data within IDOL is fully self-organizing and does not require the manual definition of complex schemas.

Parametric Probabilistic Space Analysis (PPSA)

IDOL incorporates advanced pattern recognition technologies for structured data, enabling computers to replicate the human ability to intelligently recognize and understand complex patterns in data. PPSA is a highly sophisticated parametric search capability that relates n-dimensional structured objects to one another conceptually, even where no direct field match exists.

Business Applications

Autonomy fully supports querying and integrating results from structured content residing in business applications. Autonomy provides flexible and open indexing APIs that not only import content, but also preserve and intelligently process all object rules, business intelligence and complex metadata relationships that reside in these applications. By preserving all business logic that the client has built into Line of Business (LOB) applications, it is able to interrelate between multiple databases and create a unified view of all relevant data from disparate systems.

Support for Semi-Structured Content

IDOL natively ingests XML files and fully supports the searching, processing, and analyzing of semi-structured content. Standard Boolean operators can be used to help establish relevancy, such as WHEN (structural match), WHENn (nested structural match), and vWHEN (structural weighted search), and as in structured data queries, many other search operators are also supported. IDOL allows organizations to eliminate the inefficiencies of the manual issues associated with creating XML tags by understanding the content and purpose of either the tag itself, related information, or both. Its key benefits include:

- Removing the need to manually insert XML tags
- Allowing interoperability between applications that use different XML tagging rules
- Allowing applications to use idea distancing (vital relationship between seemingly separately tagged subjects) to increase findability of information
- Automating processes that were previously performed manually
- Natively indexing XML directly into the engine
- Accessibility by XQuery as a query language
- Obtaining all output from the engine in XML format

Unified Information Access By storing all content—structured, semi-structured, unstructured, transactional and archived—in a single IDOL index, users are given a unified, holistic view of the entire enterprise knowledgebase and can realize relationships that lead to increased productivity, reduction in duplicate work, and other significant cost-saving benefits. This unified architecture enables automatic and rapid linking of information to be formed between all formats, multimedia, records and many others.

"The company's technology enables customers to decipher information from multiple sources, giving it a world-leading role"

—The Times, May 2009

Autonomy's Technology

A Different Approach

More than 80% of all data in an enterprise is unstructured information. This encompasses telephone conversations, voicemails, emails, word documents, paper documents, images, web pages, video and hundreds of other formats. Unfortunately, attempts to leverage this immense and strategic resource often fail because many businesses lack the requisite technology to understand and effectively utilize content that resides outside the scope of structured databases. Similarly, unstructured processes are equally unwieldy yet comprise the bulk of business operations. Current trends anticipate the rapid proliferation of rich media, widespread adoption of VOIP, growing use of IPTV and increased scrutiny of white-collar crimes. This overwhelming growth demands an automated solution that can effectively manage an unstructured digital morass. These concerns necessitate an information infrastructure platform that addresses all classes of information in a manner analogous to well established methods for structured databases. Akin to the Relational Database Management System (RDBMS) that revolutionized the computing industry in the 1960s, this innovative platform would enable computers to process not only structured data, but also vast amounts of semi-structured and unstructured information using a global relational index. Autonomy's ability to process all forms of digital information on a single platform offers a unique solution to a growing number of applications and devices that are increasingly dependent on utilizing unstructured information. Autonomy employs a unique combination of technologies to enable computers to form a contextual understanding of all digital content, as well as understand people's interaction with the data. Autonomy's technology eliminates the traditionally manual and costly operation of processing and analyzing information by performing these functions automatically and in real-time. This represents substantial savings for every type of organization and industry and is driving the accelerated adoption of Autonomy's technology across a diverse range of vertical markets.

"Autonomy is the market leader by a mile"

—Daily Telegraph, 2007

A Unique Combination of Technologies

Open Philosophy

Autonomy maintains an open philosophy with regards to the techniques it uses and is dedicated to selecting methods which optimize its technology, whether they are old or new. Autonomy embraces traditional or legacy methods such as keyword, Boolean, parametric and others. However, Autonomy is best known for its pioneering work in conceptual search based on computational pattern recognition (non-linear adaptive digital signal processing) and contextual linguistic analysis. Built upon the seminal mathematical works of Thomas Bayes and Claude Shannon, and on a range of innovations that are covered by 170 patents, Autonomy technology identifies the patterns that naturally occur in text, voice or video files based on the usage and frequency of terms that correspond to specific concepts. By studying the preponderance of one pattern over another, Autonomy's technology understands that there is X% probability that the content in question deals with a specific subject. In this way, Autonomy extracts the content's digital essence, encodes the unique "signature" of the concepts, and enables a host of operations to be automatically performed on emails, phone conversations, video, documents and even people's interests.

Bayesian Inference

$$P(\theta \mid x) = \frac{P(x \mid \theta) \cdot P(\theta)}{\Sigma_{\theta' \in \Theta} P(x \mid \theta') \cdot P(\theta')}$$

Thomas Bayes was an 18th century English cleric whose work has become a central tenet of modern statistical probability modeling. Bayes' efforts centered on calculating the probabilistic relationships between multiple variables and determining the extent to which these relationships are affected when new information is obtained. A traditional statistical argument posits that if a coin is tossed 100 times and comes up heads every time, it still has an even chance of coming up tails on the next throw. An alternative, Bayesian approach, is to say that 100 consecutive heads are evidence that the coin is biased. What Bayes theorem clearly demonstrated is that: a) the more information given, the more accurate the view of the world will be, and b) prior experience should be used to inform new data. In a typical problem such as judging the relevance of content to a given query, Bayesian theory dictates that this calculation be related to details that are already known. A good example of this theory at work is Autonomy's agent profile technology. Users can create agents to automatically track the latest information related to their interests, and IDOL determines the relevance of a document based on the model of the agent. Adaptive Probabilistic Concept Modeling (APCM) algorithms are also used to analyze, sort and cross-reference unstructured information. In a similar manner, knowledge about the documents deemed relevant by a user to an agent's profile can be used in judging the relevance of future documents. While most other models start with a prior knowledge of the state of the system and apply training to it, Autonomy begins with a blank slate and allows incoming data to dictate the model. In true Bayesian fashion, the model mixes new information with a growing body of older content to refine and retrain the engine.

Shannon's Information Theory

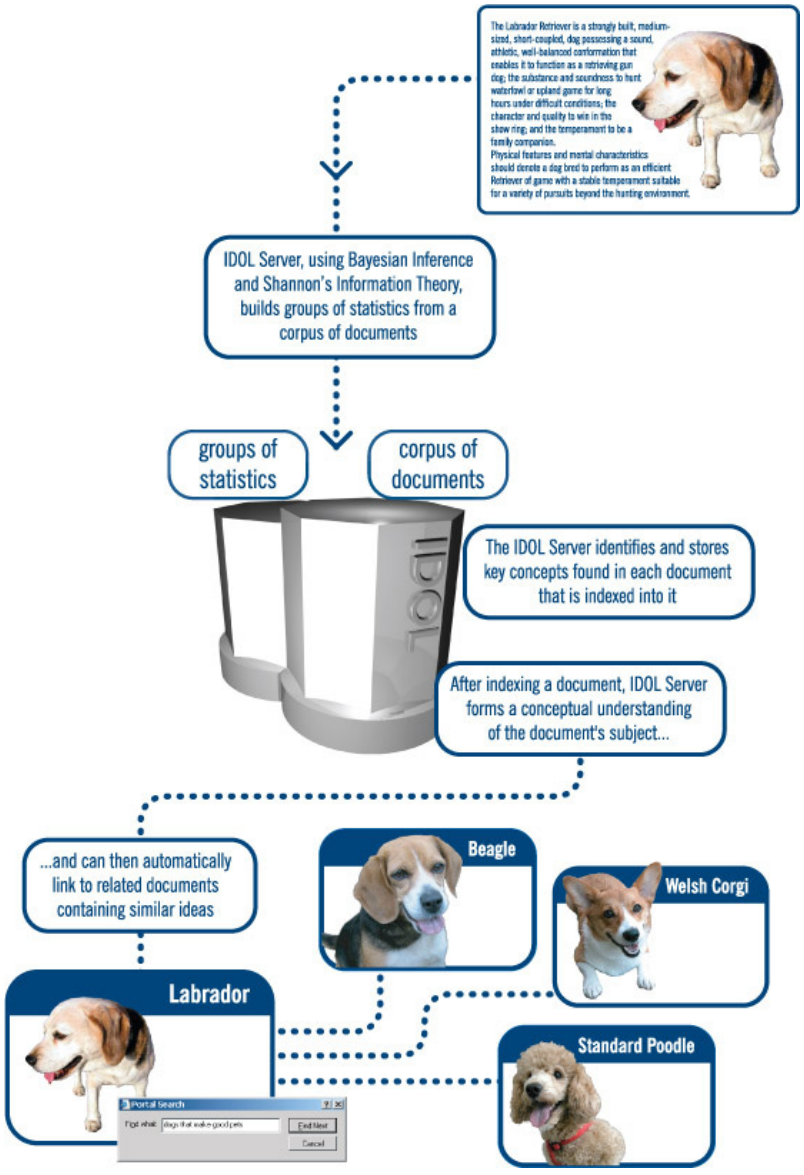
$$H = -\Sigma p_i \cdot \log_2(p_i)$$

Shannon's Information Theory forms the mathematical foundation for all digital communications systems. Claude Shannon stated that information could be treated as a quantifiable value in communications. Natural languages contain a high degree of redundancy or nonessential content. For example, a conversation in a noisy room can be understood even when some of the words cannot be heard, and the essence of a

news article can be grasped simply by skimming over the text. Information Theory provides a framework for extracting the concepts from this redundancy. Autonomy's approach to concept modeling relies on Shannon's theory that the less frequently a unit of communication occurs, the more information it conveys. Therefore, ideas, which are rarer within the context of a communication, tend to be more indicative of its meaning. It is this theory that enables Autonomy's software to determine the most important, or informative, concepts within a document.

Performance of IDOL's conceptual retrieval

Built on a unique pattern-matching technology, IDOL's conceptual query mechanism allows a seemingly simple query expression to be evaluated in complex ways; as well as the matching of the basic terms within documents using patented weighting algorithms, it is able to develop the terms to "read between the lines" and determine conceptual matches that legacy search engines would be unable to locate. However, IDOL is able to perform these evaluations with surprisingly little overhead above the equivalent keyword query. The reasons for this are two-fold. Firstly, the majority of the work in the calculation and initialization of the conceptual matching is done at index time, as opposed to query time; the documents are analyzed while the data is being processed to form a statistical "pool" from which queries can draw key conceptual information, as well as an overlying Bayesian network in which apparently unrelated pieces of information are automatically linked via dynamic probabilities. The second reason is that the document-matching algorithm itself within IDOL uses widespread "short-circuiting" and iterative calculation to ensure that it only performs exactly as much calculation as is required. In essence, the key conceptual information is already available before the query has even started, and once it does begin, it feeds directly from the statistical core to load the information. The uniqueness of the query then forces the only truly complex step, a one-off calculation in which combination algorithms arrive at the most relevant set of documents to the query. These can then be returned without the need for looping through every potential match.



Manual or Automatic - It's Not an Either/Or Choice

Avoiding Black Box Solution Pitfalls

Some vendors only offer "black box" solutions, mistakenly believing that their technology can always provide the best answers with no tuning required. However, this idea demonstrates a naïve understanding of enterprise demands, for not even the best of automated systems can anticipate the special needs of each enterprise. These "black boxes" offer only a few, if any, tuning options for relevancy and do not reveal how the results were generated. In stark contrast, Autonomy's technology provides the best of both worlds, automatically retrieving the most accurate results using its conceptual understanding of content and also offering the flexibility to modify the relevancy algorithm if needed. The computational process is fully transparent to the administrator and Autonomy reveals the basis for its determinations through easily understood representations such as dominant terms and idea distances. Covered in more detail in the Administration section of this book, both system administrators and business users are provided with a full workbench to control and tune the relevancy of search results. Some unique advantages offered by Autonomy include:

- WYSIWYG (What You See Is What You Get) user interface
- The weight of virtually every field (e.g. title, author) can be manipulated and many operators are available to alter relevancy
- Full support for voting and document rating. End users can rate the usefulness of specific documents, and this information is subsequently used to calculate relevancy or primacy. The voting can be limited to those people whose profiles show a strong match to the subject of the document
- Access to an extensive range of pertinent information, including common queries, misspellings, query types – all presented in friendly visuals
- Full support for business modulated result "sponsoring" or "placement." For example, a business user can elect to promote a certain result, set of results or object (such as an advertisement) to a defined position within a result set in response to a given query or input. If a user queries for "yellow Prius," the administrator can define a rule to return the same set of results as a query for "gold Prius," with a link to the advantages of hybrid cars being the first on the returned list
- Autonomy Collaborative Classifier module, which creates a workflow in which the subject matter experts and knowledge engineers, as identified by the organization, collaborate in real-time to create, modify, distribute and manage taxonomies. As the classifications are created and managed by the people who actually use them, information is organized in ways that are specific and germane to the organization
- Protects user privacy by respecting entitlement rights and separating the administrator role from the "super-user", thereby ensuring the administrator will be restricted from information they are not privileged to view

In addition to providing administrators with comprehensive tools to alter the relevancy modeling, Autonomy is transparent in the methods it uses to arrive at such results. Autonomy uses the full text of the document to determine relevancy, and even

with no manual configuration, administrators and users can easily understand how the results were selected. Autonomy uses many ways to justify relevancy, these include:

- **% relevancy:** Provides percentage similarity of the document to the query
- **Automatic highlighting:** Highlights key terms/concepts within the document
- **Ideas cloud:** Presents list of concepts present in the results list, with variable font size and boldness to represent the number of results that include that concept
- **Cluster tree:** Displays hierarchy of relevant entities (concepts or metadata) extracted from the results list; it displays the count of documents that contain that entity
- **Automatic summarization of content:** Demonstrates the key concepts extracted, which may come from different parts of the document
- **Query journey:** Delineates the logical path that IDOL took to arrive at a given set of results by revealing the key concept terms that were found, along with pertinent metadata, repositories searched and the relevancy threshold reached
- **Query bread-crumbing:** Traces all aspects of a user's query interaction

Autonomy enables an entire range of information processing options, both manual and automatic. The system can be configured to support as much or as little manual involvement as necessary, ensuring that Autonomy is not a "black box" where the running of the technology cannot be seen or adapted by administrators

Information Connectivity

Currently connecting to over 400 content repositories and supporting over 1,000 file formats, Autonomy is uniquely able to aggregate and index any form of structured, semi-structured and unstructured data into a single index, regardless of where the file resides. In today's enterprise environment where most users typically waste time searching in four or five repositories to arrive at an acceptable results list, Autonomy's extensive set of connectors enables a single point of search for all enterprise information (including rich media), saving organizations much time and money. With access to virtually every piece of content, IDOL provides a 360 degree view of an organization's data assets. Additionally Autonomy allows you to develop connectors to any obsolete or bespoke system no longer supported by other vendors, in addition to allowing you to develop connectors for your own future repositories. This ensures that your past investments are safeguarded whilst enabling you to adopt new systems confident in the knowledge that they will be fully supported by Autonomy. As Autonomy wholly owns all its software, including connectors, you will always be in the position where you can renew the license no matter what other changes go on in the broader software sector.

"Autonomy's unique Meaning-Based platform enables organizations to seamlessly incorporate untapped resources, such as phone recordings and emails, into their corporate strategy and benefit from a single point of access to all of their information."

—Keith Dawson, Research Analyst, Frost & Sullivan

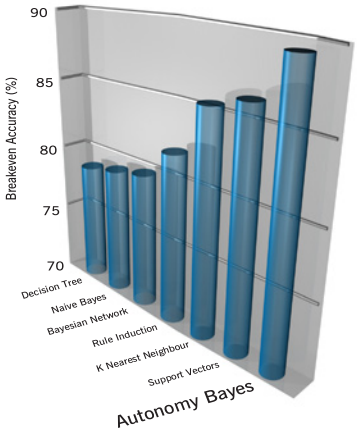
Differentiation

Autonomy offers a single platform in IDOL for understanding the meaning of all human-friendly information, including emails, web pages, social media, blogs, audio and video, as well as people's interactions with that data, enabling content and interactions to be processed more intelligently. Only IDOL can perform keyword and conceptual search, speech analytics, video search, email and IM search, and categorization – all on the same platform. Alternative approaches require the stitching together of different technologies with potentially conflicting formats. Consequently, they compromise stability, pose maintenance issues and may necessitate excessive technical support when upgraded. IDOL, as a single platform, circumvents these problems, and its benefits are realized without any concessions in performance; as each function of IDOL is best of breed.

IDOL is data agnostic, language independent and fully scalable. It does not require complex programming, extensive integration, business rules or middleware. It also does not require information to be manually tagged, linked or categorized. IDOL-compliant applications are immediately compatible through their common understanding of digital information. All this is possible because IDOL understands information in a manner similar to humans – it directly relates concepts "read" from the portions of digital content that humans can process - not from rules that are dependent on synthetic tags.

Accuracy

Autonomy's pioneering MBC platform provides advanced and accurate enterprise retrieval by encompassing and transcending existing search mechanisms. By searching using the concepts defined by the given words, and not just keyword techniques, it provides superior accuracy and retrieves the most relevant content, not just the most popular. The user can further enhance their findings upon query by applying one of the many content operations available.



Moreover, enabled by its modular architecture and rapid indexing rate, IDOL extracts and utilizes all of the content within a file instead of just the first few pages. Other technology vendors use partial indexing to provide the impression of fast performance. IDOL indexes all of the content of a file along with its metadata and concepts, ensuring rapid and precise searching. IDOL presents a complete set of search results with no premature cessation jump out, enabling full discovery. This is especially crucial in legal matters, where searches that quit prematurely after retrieving an arbitrary number of results can have disastrous consequences.

Automation

By automating processes that were previously performed by costly and tedious labor, Autonomy's technology offers a direct path to substantial bottom line savings. Cross-referencing of content is automatic as Autonomy's infrastructure identifies related material within the operating layer and determines relationships between information using multi-tiered relevancy modeling. Processes such as hyperlinking, information clustering, alerting and categorization of content can all be precisely automated with any document or set of documents.

Cross-lingual Functionality

Autonomy's technology allows for cross-lingual search and data management. There is no compromise to the accuracy and concepts extracted regardless of the language used. For example, an employee based in New York and working in English can be provided with relevant documents and information from all over his company's global network, whatever language they are in, based on the similarity of concepts expressed within the documents. This language agnostic approach offers a significant benefit to global businesses by enabling colleagues separated by distance and language to collaborate and share knowledge.

Language Independence

Autonomy's technology uses probabilistic modeling to extract meaning from content, and forgoes language-dependent parsing or dictionaries to form ideas. Because Autonomy treats words merely as abstract symbols of meaning, it is completely language independent. It does not rely on an intimate knowledge of a language's grammatical structure, but rather derives its understanding through the context of the words' occurrence. This sophisticated technique yields high accuracy and performance is further optimized through proprietary stemming algorithms, "sentence-breaking" libraries, stoplists and n-grams.

bbcarabic.com

Although IDOL is fundamentally predicated on a language independent model, it is still capable of using linguistic analysis to parse semantics to an intra-document level. For instance, the Sentiment Analysis functionality can determine the degree to which a sentiment is positive, negative or neutral for the entire content or a segment of the content. For example, a blogger may have a positive opinion on the iPod, but a negative one on the iPhone, all within the same entry. In addition, by extracting information from every file processed, IDOL continually learns positive and negative language structures and concepts.

IDOL supports over 100 languages, including English, German, French, Italian, Chinese and Japanese, and can be easily configured to auto-detect the language of incoming documents.

Learning Ability

Due to the unique combination of Bayesian Inference and Shannon's Information Theory at the core of the technology, Autonomy software is able to continuously develop and learn. This learning ability significantly reduces the manual input required by other solutions and translates into large savings in time and money. Whereas other solutions need to be taught new words or phrases and shown how to categorize them, Autonomy can automatically deduce the significance of these new units of meaning, add them to relevant categories and create new categories where necessary. Autonomy's technology can also learn about its users by dynamically monitoring the content they view, and then deliver new and relevant content as it is added to the environment.

Format Agnosticism

Autonomy is not reliant on any single file or data format. Autonomy aggregates more than 1,000 content formats, including voice and video content, from the most comprehensive range of repositories. As a result, Autonomy allows enterprises to make sense of information from the widest range of formats and sources available, from unstructured data (e.g. HTML pages, word processing documents, spreadsheets, email, multimedia content) to semi-structured (e.g. XML) and structured data (e.g. Oracle, Lotus Notes, ODBC compliant material). In addition, Autonomy penetrates the information silos in an enterprise by offering deep integration into a myriad of repositories including Documentum, SharePoint, Lotus Notes, Exchange, RDBMS, file servers and more.

Innovation: 170 Patents

As a groundbreaking company founded out of pioneering research at Cambridge University, Autonomy is proud to maintain its reputation as one of the most innovative companies in the world. Autonomy continues to focus on research and development (R&D) to deliver increasingly innovative products to customers and consolidate its position as a market leader. Autonomy's commitment to R&D is evident in the company's continued investment in this area, with R&D investment increasing every year.

Autonomy owns 100% of its technology and hence eliminates the uncertainty of sourcing and integrating solutions from multiple vendors in a consolidated market space. No matter what changes take place in the broader market, Autonomy customers will never be left with unsupported technology or software licenses that become non-renewable. As the largest department in the group, the R&D team continues to represent the lifeblood of the company.

In recent years, Autonomy has successfully launched a number of new solutions as a result of extensive R&D, introducing industry-shifting technology into the marketplace. As a testimony to unsurpassed innovation, Autonomy is quickly gaining mainstream adoption and is being established as the standard for all information needs for organizations around the globe.

Support for Future Technologies

Autonomy’s core meaning-based technology addresses such fundamental problems that its solutions are being used in virtually every market, including situations that would have been unimaginable when the technology was conceived ten years ago. From handling the latest rich media content to providing eDiscovery support for current SEC compliance regulations, Autonomy has demonstrated its extensibility by growing and adapting to new requirements and technologies. Autonomy's Products

Autonomy delivers a complete software infrastructure solution that forms an understanding of content in any file format - text or voice-based, structured or unstructured - regardless of where it is stored, how it was created, or which application is associated with the data. Applications can thereby communicate with each other without any manual effort involved in setting up complicated connectors or using metadata. Autonomy's technology makes enterprise systems "data-agnostic" and provides automated efficiencies for processes that were once manually intensive. Autonomy's unique infrastructure technology solves a fundamental problem that affects every industry, as evidenced by its almost universal deployment in over 20,000 customers worldwide and embedding in over 400 OEMs. Customers can leverage content and drive interactions from existing applications that handle unstructured information, including enterprise search projects, eCommerce, KM and e-learning portals, mobile devices, Business Intelligence, CRM, BPM, IRM, consolidated archiving and eDiscovery, rich media management, and security and surveillance.

Independent analysts such as Gartner, Butler Group, Delphi Group, AMR Research, Forrester Research and Ovum, as well as news outlets, continue to praise Autonomy for its ongoing development and innovative products.

"Queen's Awards for Enterprise: Innovation"
Autonomy IDOL, 2009

"A Top 100 Company in the Digital Content Industry"
Autonomy - eContent Magazine

"2009 Editors' Choice Award"
Autonomy - Intelligent Enterprise

"100 Companies that Matter in Knowledge Management"
Autonomy -2009 KMWorld Magazine

"Entrepreneur of the Year"
Autonomy CEO Dr. Mike Lynch - Management Today's Top 100 Entrepreneurs 2009

"2008 Strong Performer Rating"
The Forrester Wave: Marketing Asset Management (MAM), Q1 2008

"2008 North American Customer Value Enhancement Award"
Autonomy Customer Interaction Analytics – 2008 Frost and Sullivan

"Leader in Innovation for Email and IM Compliance"
Autonomy - Financial i Magazine, Leaders in Innovation Awards 2008

"Company of the Year"
Autonomy - techMARK Awards 2008

"Badenoch and Clark Business of the Year"
Autonomy - National Business Awards 2008

"Top 5 Electronic Discovery Provider"
Autonomy - 2008 Socha-Gelbmann Electronic Discovery Survey Report

"2008 Trendsetter Award"
Autonomy IDOL - KMWorld Magazine

"2008 CRM Excellence Award"
Autonomy Intelligent Contact Center - Customer Interaction Solutions Magazine

"Autonomy CEO and founder Dr Mike Lynch Named "Innovator of the Year"
Autonomy - 2008 European Business Leaders Awards

"Best Performing Software Company in Europe"
Autonomy - 2008 Truffle 100

"Best Government Solution and Best Technology Provider"
Autonomy - 2007 Gartner IT ChannelVision

"Top 100 Fastest Growing Companies"
Autonomy - 2007 Deloitte

"Award for Excellence in Technology"
Autonomy Virage - 2007 Frost & Sullivan Award

(Autonomy Inc. and Autonomy Systems Limited are both subsidiaries of Autonomy Corporation plc.)